Taking a Look at Small-Scale Pedestrians and Occluded Pedestrians

Jiale Cao, Yanwei Pang, *Senior Member, IEEE*, Jungong Han, *Senior Member, IEEE*, Bolin Gao, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Small-scale pedestrian detection and occluded pedestrian detection are two challenging tasks. However, most state-of-the-art methods merely handle one single task each time, thus giving rise to relatively poor performance when the two tasks, in practise, are required simultaneously. In this paper, it is found that small-scale pedestrian detection and occluded pedestrian detection actually have a common problem, i.e., inaccurate location problem. Therefore, solving this problem enables to improve the performance of both tasks. To this end, we pay more attention to the predicted bounding box with worse location precision and extract more contextual information around objects, where two modules (i.e., location bootstrap and semantic transition) are respectively proposed. The location bootstrap is used to re-weight the regression loss, where the loss of predicted bounding box far from the corresponding ground-truth is up-weighted and the loss of predicted bounding box near the corresponding ground-truth is down-weighted. Meanwhile, the semantic transition adds more contextual information and relieves the semantic inconsistency of skip-layer fusion. Since the location bootstrap is not used at the test stage and the semantic transition is light-weight, the proposed method does not add much extra computational costs during inference. Experiments on the challenging Citypersons and Caltech datasets show that the proposed method outperforms the state-of-the-art methods on the small-scale pedestrians and occluded pedestrians (e.g., 5.20% and 4.73% improvements on the Caltech).**

***Index Terms*—Small-scale pedestrians, occluded pedestrians, location bootstrap, and semantic transition.**

I. Introduction

Pedestrian detection aims to classify and locate pedestrians in a given image, which can be applied to self-driving cars, human-computer interaction, video surveillance, *etc*. In recent few years, pedestrian detection based on deep convolutional neural networks has achieved immense progress [22], [18], [43], [26], [45]. Despite the great success, small-scale pedestrian detection and occluded pedestrian detection are still very challenging. Thus, improving detection performance of small-scale pedestrians and occluded pedestrians is very important and necessary.

To solve small scale or occlusion problems on pedestrian detection, many attempts have been made by the researchers in the past few years. The related methods can be summarized as follows: (1) *small scale aspect*. Some methods treat the objects at different scales as different pedestrian subcategories and detect them separately [23], [51], [5]. Some other methods use the features of large-scale objects to guide the feature learning of small-scale objects [28], [25], [40]. (2) *occlusion aspect*. Most methods integrate some deep part models to improve occluded pedestrian detection [47], [56], [56]. Alternatively, Repulsion Loss [49] gathers the proposals belonging to the same object for crowded pedestrian detection.

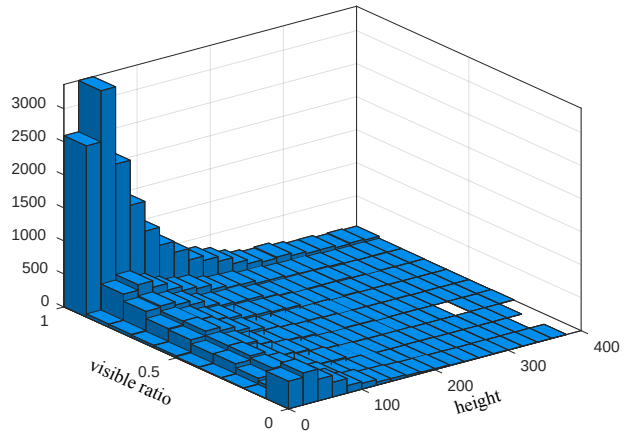


Fig. 1. The histogram of pedestrian height and visible ratio. It can be seen that many occluded pedestrians are under 100 pixels tall, which can be seen as small-scale pedestrians. It means that small-scale pedestrians and occluded pedestrians are related in some degree.

However, these methods merely focus on either small-scale problem or occlusion problem on pedestrian detection. We argue that small-scale pedestrians and occluded pedestrians are related in some degree. Fig. 1 plots the histogram of pedestrian height and visible ratio. It can be seen that many occluded pedestrians belong to small-scale pedestrians. Meanwhile, small-scale pedestrians and occluded pedestrians both lack visual information in some degree such that accurately localizing them becomes problematic. When the evaluation IoU threshold on Citypersons [54] varies from 0.5 to 0.3, we find that the miss rates of FPN on the small-scale subset and occlusion subset of Citypersons have 8.8% and 7.6% drops. It means that many small-scale pedestrians and occluded pedestrians can be found but not accurately located, which is called *inaccurate location problem*.

To solve the above problem, it is believed that more attention to the predicted bounding boxes with worse location

J. Cao and Y. Pang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (E-mail: {connor,pyw}@tju.edu.cn).

J. Han is with the School of Computing and Communications, Lancaster University, UK (E-mail: jungong.han@lancaster.ac.uk)

B. Gao is with the China Automotive Technology and Research Center Co., Ltd., Tianjin 300300, China (E-mail: gaobolin@catarc.ac.cn).

X. Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: li@nwpu.edu.cn).

precision and more contextual information extracted around objects are useful. As a result, we propose two modules, i.e., location bootstrap and semantic transition, for both small-scale and occluded pedestrian detections. The location bootstrap pays more attention to the predicted bounding boxes having relatively worse location precision. Specifically, the averaged IoU between bounding boxes and their corresponding ground-truths is calculated at each training iteration. After that, the loss of bounding box which has the IoU over the averaged IoU is down-weighted, while the loss of bounding box having the IoU below the averaged IoU is up-weighted. The semantic transition model uses the large kernel convolution as the lateral connection of skip-layer fusion to extract more contextual information and relieve the semantic inconsistency of skip-layer fusion. The contributions and the metrics of this paper are summarized as follows:

(1) We point out an inspiring characteristic of small-scale pedestrian detection and occluded pedestrian detection. That is, many small-scale pedestrians and occluded pedestrians can be found but not accurately located.

(2) Inspired by the above characteristic, the location bootstrap and semantic transition modules are proposed. Location bootstrap down-weights the regression loss of bounding box with high location precision and up-weights the loss of bounding box with low location precision. Semantic transition module uses the large kernel convoltion to extract more contextual information and relieves the semantic inconsistency.

(3) Experiments on the Citypersons [54] and Caltech [15] datasets show the effectiveness of proposed methods. Meanwhile, the proposed method does not add much computational costs.

The rest of this paper is organized as follows. The related works are given in Section II. After that, the proposed method is introduced in Section III. Then, experiments are shown in Section IV. Finally, Section V concludes this paper.

II. Related works

With the success of deep Convolutional Neural Networks (CNN), pedestrian detection has achieved great progress in recent few years. In this section, a review of deep pedestrian detection is firstly given. Afterwards, we focus on explaining the improvements on small-scale pedestrian detection and occluded pedestrian detection particularly.

A. A review of pedestrian detection

Pedestrian detection can be divided into two main classes: handcrafted features based methods and CNN based methods. Usually, handcrafted features based methods use the handcrafted features and the shallow classifiers (e.g., AdaBoost and SVM) to learn pedestrian detectors. In 2009, Dallár *et al.* [14] proposed Integral Channel Features (ICF), which firstly converts the original RGB image to ten feature channels (i.e., histogram of gradients, gradient magnitude, and LUV color channels) and secondly uses Cascade AdaBoost to learn the detector based on the features extracted from the ten channels. Based on ICF [14], many variants including the local features (e.g., ACF [13], Checkerboards [55], and LDCF [38]) and the non-local features (e.g., NNNF [7]) have been proposed.

In the recent few years, CNN based methods have achieved great success on object detection and pedestrian detection [53], [4], [31], [59], [1], [2], [3]. At first, CNN [26], [45] is simply acted as the feature extractor for pedestrian detection, which is fed to the shallow classifier. For example, Yang *et al.* [50] proposed to use the convolutional channel features to replace the filtered channel features. Cao *et al.* [8] integrated the handcrafted feature channels and each layer of CNN into multilayer feature channels. With the success of Faster RCNN [43] on general object detection, many end-to-end CNN based methods have been also proposed on pedestrian detection. For example, Zhang *et al.* [54] proposed AdaptedRCNN by modifying some network settings to better detect pedestrians. Mao *et al.* [37] proposed to add semantic segmentation task to help improve pedestrian detection. Recently, some one-stage methods (e.g., ALFNet [36], GDFL [27], and OHNH [39]) have been also proposed for pedestrian detection.

Contextual information is important for object detection and pedestrian detection. Zeng *et al.* [52] proposed a gated bi-directional CNN to adaptively model the interactions of contextual and local visual cues. Wang *et al.* [48] proposed to use two different branches to respectively extract the body semantic and contextual information for pedestrian detection. Li *et al.* [30] proposed a novel attention module to exploit the better context. Li *et al.* [29] proposed a generic context-mining RoI operator to extract good contextual information around the proposals. To refine the proposals, Chen *et al.* [10] proposed a contextual refinement module to aggregate the rich contextual information. These methods mainly extract the contextual information of ROI proposals, but do not extract context of the feature maps so that the semantic gap of feature pyramid network [32] is still there. Contextual information is also useful in segmentation. For example, the large kernel and dilated convolutions [41], [9] are proposed to extract multi-scale context.

B. Small-scale pedestrian detection

Compared with large-scale pedestrians, small-scale pedestrians are relatively blurred and noisy. To improve the detection performance of small-scale pedestrians, many attempts have been made by the researchers. Some researchers proposed to respectively detect the pedestrians at different scales. For example, Li *et al.* [23] proposed to use two subnetworks to respectively detect small-scale pedestrians and large-scale pedestrians. Similarly, Yang *et al.* [51] proposed scale dependent ROI pooling for multi-scale object detection. Cai *et al.* [5] proposed to use the features of different layers to detect the objects at different scales. To enhance the feature semantics of different output layers, Lin *et al.* [32] proposed to combine the weak semantic features with the strong semantic features by top-down structure. Among the single-stage methods, SSD [35] and YOLOv3 [42] use the layers of different spatial resolutions to detect objects at different scales. DSSD [17] uses the top-down structure to enhance the features of SSD. Meanwhile, some researchers proposed to narrow the difference between the features of pedestrians at different scales. For

Fig. 2. Some examples of inaccurate location on small-scale pedestrians or occluded pedestrians. The number in the image means the overlap between detected bounding box (blue) and ground-truth (red). Namely, many small-scale pedestrians and occluded pedestrians can be found but not well located.

example, based on the technique of Generative Adversarial Networks (GAN) [19], Li *et al.* [28] proposed to learn the residual between the features of small-scale objects and the features of large-scale objects. Kim *et al.* [25] proposed a scale aware network, which maps the features at different scales into a scale-invariant subspace. Except for these methods, Song *et al.* [46] proposed to detect small-scale pedestrians by using somatic topological line localization.

C. Occluded pedestrian detection

Occluded pedestrian detection is also challenging due to the visual information loss in object parts. To solve occlusion problem, the researchers proposed to pay more attention to the visible parts of objects [47], [56], [57], [39]. For example, Tian *et al.* [47] proposed to train multiple deep part detectors and integrate their detection scores together by the linear SVM. Zhang *et al.* [56] proposed a part occlusion-aware region of interest (PORoI) pooling to detect occluded pedestrians. Zhou *et al.* [60] proposed to use the visible part label to reduce the effect of occlusion and help improve occluded pedestrian detection. Zhang *et al.* [57] proposed to apply the channel-wise attention to handle different occlusion patterns for pedestrian detection. Alternatively, Wang *et al.* [49] proposed a repulsion loss which can better gather the candidate proposals belonging to the same object in the crowded scenes.

III. The Proposed Method

In this section, we give a detailed description about the proposed method. Firstly of all, the motivation and the overall architecture of proposed method are given. After that, two main modules (i.e., location bootstrap module and semantic transition module) are described in detail.

A. Overview

Motivation: Small-scale pedestrian detection and occluded pedestrian detection are two challenging tasks. We argue that small-scale pedestrians and occluded pedestrians are very related. Based on Fig. 1, it can be seen that many occluded pedestrians belong to small-scale pedestrians. As a result, solving the common problem of small-scale pedestrians and occluded pedestrians can largely improve detection performance of pedestrian detection.

In fact, the visual information on both small-scale pedestrians and occluded pedestrians is not as adequate as the normal pedestrians, so that it is difficult to accurately detect them. By loosening the evaluation metric between bounding boxes and ground-truth from 0.5 to 0.3, the detection performance of many small-scale pedestrians and occluded pedestrians on Citypersons [54] both have about 8% improvement. It means that to some degree many small-scale pedestrians and occluded pedestrians can be found but not accurately located. For simplification, this problem is called *inaccurate location problem*. Fig. 2 further shows some examples of inaccurate location on small-scale pedestrians or occluded pedestrians.

Solution: To solve the location inaccuracy problem and improve pedestrian detection, we think that two aspects can be considered. (1) More attention should be paid to the predicted bounding boxes with relatively worse location precision, which aims to force them learn more accurate locations. Unlike cascade regression [24], [6], the proposed method does not need extra network modules. (2) More contextual information around the objects should be extracted. Because small-scale pedestrians and occluded pedestrians have limited visual information, context can provide more useful features for the better location and classification of small-scale pedestrians and occluded pedestrians. Based on the above observation and analysis, we propose two modules (i.e., location bootstrap and semantic transition) for pedestrian detection.

Architecture: As a successful architecture of object detection, Feature Pyramid Network (FPN) [32] is effective to solve scale variance problem. Thus, FPN is chosen as the basic architecture of the proposed method. The main process of FPN is introduced as follows: given an input image, it firstly goes through a backbone network (e.g., VGG16 [45] and ResNet50 [21]) to generate the feature maps of different spatial resolutions. After that, a top-down structure with the skip-layer fusion is used to generate the output feature maps which have strong semantics. Finally, object proposals are firstly generated by the multiple output feature maps and secondly classified by Fast RCNN headnetwork.

Based on FPN, the framework of proposed method is shown in Fig. 3, which incorporates two new modules (i.e., location bootstrap module and semantic transition module). The location bootstrap module is added after the ROI pooling layer, which is used to re-weight the regression loss of candidate proposals at the training stage. Specifically, it up-weights the regression loss of bounding box which has low location precision and down-weights the regression loss of bounding box which has high location precision.

To extract more contextual information around the objects, the semantic transition module uses the large kernel convolution as a lateral connection for skip-layer fusion. Meanwhile, it relieves the semantic inconsistency of two input feature maps in skip-layer fusion.

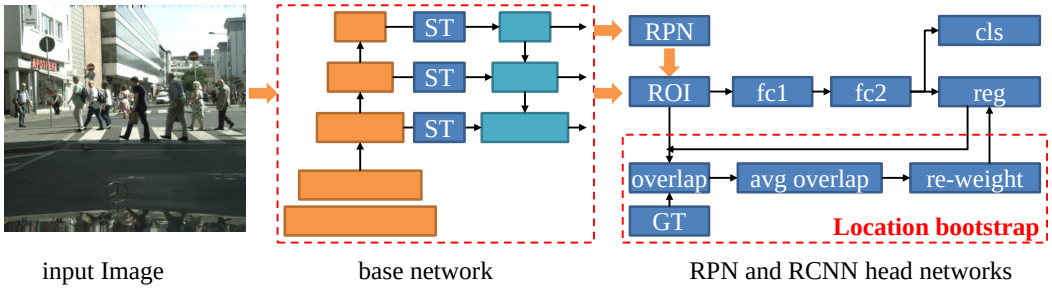


Fig. 3. The framework of proposed method for object detection based on feature pyramid network, which consists of two new modules (i.e., the location bootstrap and semantic transition module).

Algorithm 1 The process of location bootstrap.

Input:

The candidate object proposals (i.e., $B_1^o, B_2^o, ..., B_N^o$) generated by RPN, N is the number of bounding boxes;
The ROI feature map of each bounding box;
The ground-truth of each bounding box (i.e., $G_1, G_2, .., G_N$) ;

Output:

The updated weight of the regression loss for each bounding box;

1: Generate the predicted offset for each proposal B_i^o by Fast RCNN head-network with ROI feature map;
2: Calculate the predicted bounding box B_i^p for each proposal B_i^o;
3: Calculate the overlap O_i (i.e., intersection over union) between predicted bounding box B_i^p and corresponding ground-truth G_i for each proposal B_i^o;
4: Calculate the averaged overlap by the overlaps between candidate proposals and corresponding ground-truths;
5: Update the loss weight of each proposal for training.

B. Location Bootstrap Module

Because small-scale pedestrians and occluded pedestrians are much harder to be located very well, we argue that it should pay more attention to the bounding boxes with worse location precision during training. Thus, the location bootstrap module is proposed to re-weight the regression loss of bounding boxes, which up-weights the loss of the bounding box far from the corresponding ground-truth and down-weights the loss of the bounding box near the corresponding ground-truth at the training stage. As shown in Fig. 3, a location bootstrap module is added after the ROI pooling layer. Though the idea is similar to OHEM [44] and Focal Loss [34], our method is different from them. OHEM [44] and Focal Loss [34] pay more attention to the hard samples which are difficult to be well classified. Different from OHEM and Focal Loss, our method pays more attention to the samples which are difficult to be accurately regressed.

The detailed process of location bootstrap at each training iteration can be summarized in Algorithm 1. Given the candidate object proposals (i.e., B_1^o, B_2^o,..., B_N^o) generated by Region Proposal Network (RPN), Fast RCNN head-network can output the regression offsets of all the candidate proposals based on the ROI features. After that, the predicted bounding boxes of candidate proposals can be calculated. With the predicted bounding boxes B_i^p and their corresponding ground-truths G_i, the overlap O_i between the predicted bounding box B_i^p and corresponding ground-truth G_i can be calculated as follows:

$$O_i = \frac{B_i^p \bigcap G_i}{B_i^p \bigcup G_i}, i = 1, ..., N, \tag{1}$$

where N is the number of predicted candidate proposals in a mini-batch which belong to objects. The mini-batch contains the proposals of all the images at each iteration. Then, the averaged overlap O_m over a mini-batch can be written as follows:

$$O_m = \frac{1}{N} \sum_{i=1}^{N} O_i. \tag{2}$$

With the averaged overlap O_m and the overlaps between the predicted bounding boxes and corresponding ground-truths (i.e., O_1, O_2,..., O_N), the updated weight of each bounding box w_i^u can be finally calculated as

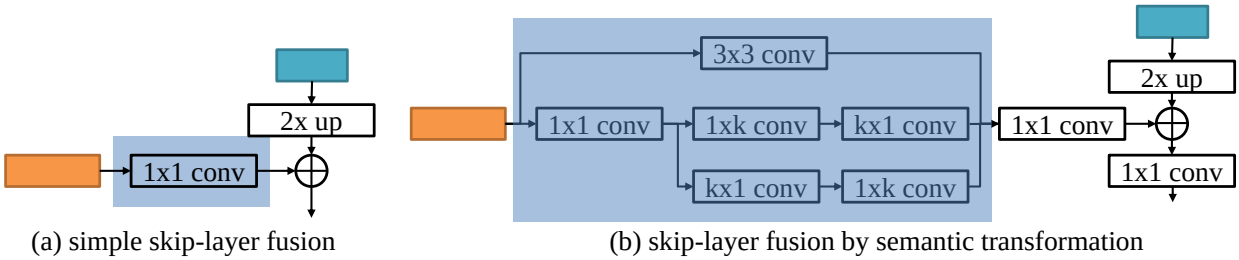
$$w_i^u = \frac{2}{1 + \exp(-\alpha \times (O_m - O_i))}, \tag{3}$$

where α is a parameter which is set as 2.0 by cross-validation. If the predicted bounding box is far from corresponding ground-truth, then $O_i < O_m$ and $w_i^u > 1$. As a result, the loss will be up-weighted. If the predicted bounding box is near the corresponding ground-truth, then $O_i > O_m$ and $w_i^u < 1$. As a result, the loss will be down-weighted.

Because the location bootstrap module is used to change the weight of the regression loss, it can be removed at the test stage. As a result, the proposed module does not increase any computational costs at the test stage.

C. Semantic Transition Module

Because small-scale pedestrians and occluded pedestrians lack some useful visual information, we think that context features around the objects become more import. Though FPN [32] incorporates some contextual information by the skip-layer fusion (see Fig. 4(a), we argue that it is still not good enough. On the one hand, using a simple 1×1 convolutional layer as skip-layer connection to combine the feature maps with weak semantics and those with strong semantics exists



(a) simple skip-layer fusion (b) skip-layer fusion by semantic transformation

Fig. 4. The structure of semantic transition module. (a) Skip-layer fusion in FPN, (b) Skip-layer fusion in our semantic transition module, which aims to extract more contextual information.

the semantic inconsistency. On the other hand, the contextual information around the objects are not adequately added.

To alleviate semantic inconsistency and enrich contextual information, semantic transition module is proposed, which uses the separable large-kernel convolution as lateral connection. Fig. 4(b) shows the skip-layer fusion by using the semantic transition module. Firstly, the feature maps of weak semantics go through three different branches. The top branch goes through a 3×3 convolutional layer. The middle and bottom branches firstly share a 1×1 convolutional layer to reduce the channel number. After that, the middle branch goes through a $1 \times k$ convolutional layer and a $1 \times k$ convolutional layer, and the bottom branch goes through a $k \times 1$ convolutional layer and a $1 \times k$ convolutional layer. In this paper, $k = 7$. Finally, the feature maps generated by three branches are concatenated together and fed to a 1×1 convolutional layer for final skip-layer fusion. The top branch of 3×3 convolution aims to extract local features and two bottom branches of large kernel convolutions are used to extract more contextual information. To reduce computational cost and network parameters, separable large kernel covolution is used.

Compared to the simple skip-layer fusion in Fig. 4(a) of FPN, the proposed semantic transform model in Fig. 4(b) can relieve the semantic inconsistency of skip-layer fusion and add more contextual information. Meanwhile, with the separable convolution, it does not increase much computational cost.

IV. Experiments

In this section, some experiments on two famous pedestrian datasets (i.e., the Citypersons dataset [54] and the Caltech pedestrian dataset [15]) are conducted to demonstrate the effectiveness of proposed methods and compare with some state-of-the-art methods.

A. Datasets and Evaluation

The Citypersons dataset [54] is an extended pedestrian dataset by using the instance information of the Cityscapes benchmark [11], which consists of three subsets (i.e., `trainval`, `val`, and `test`). The `trainval` set has 2975 images for training, the `val` set has 500 images for ablation study, and the `test` set has 1525 images for performance evaluation.

The Caltech pedestrian dataset [15] is a very famous pedestrian dataset, which is captured on the vehicle car in the urban street. It consists of 11 videos, where the first 6 videos are used for training and the last 5 videos are used for test. To enlarge the training data, the training images are densely sampled per three frames from the training videos. Thus, there are 42782 images for training. The standard test images are sampled per thirty frames from the test videos. Thus, there are 4024 images for performance evaluation.

TABLE I
Ablation experiments on Citypersons. **R** means the `reasonable` set, **RS** means the `small` set, **HO** means the `heavy` occlusion set, **R+HO** means the `reasonable+heavy` set, and **A** means the `all` set.

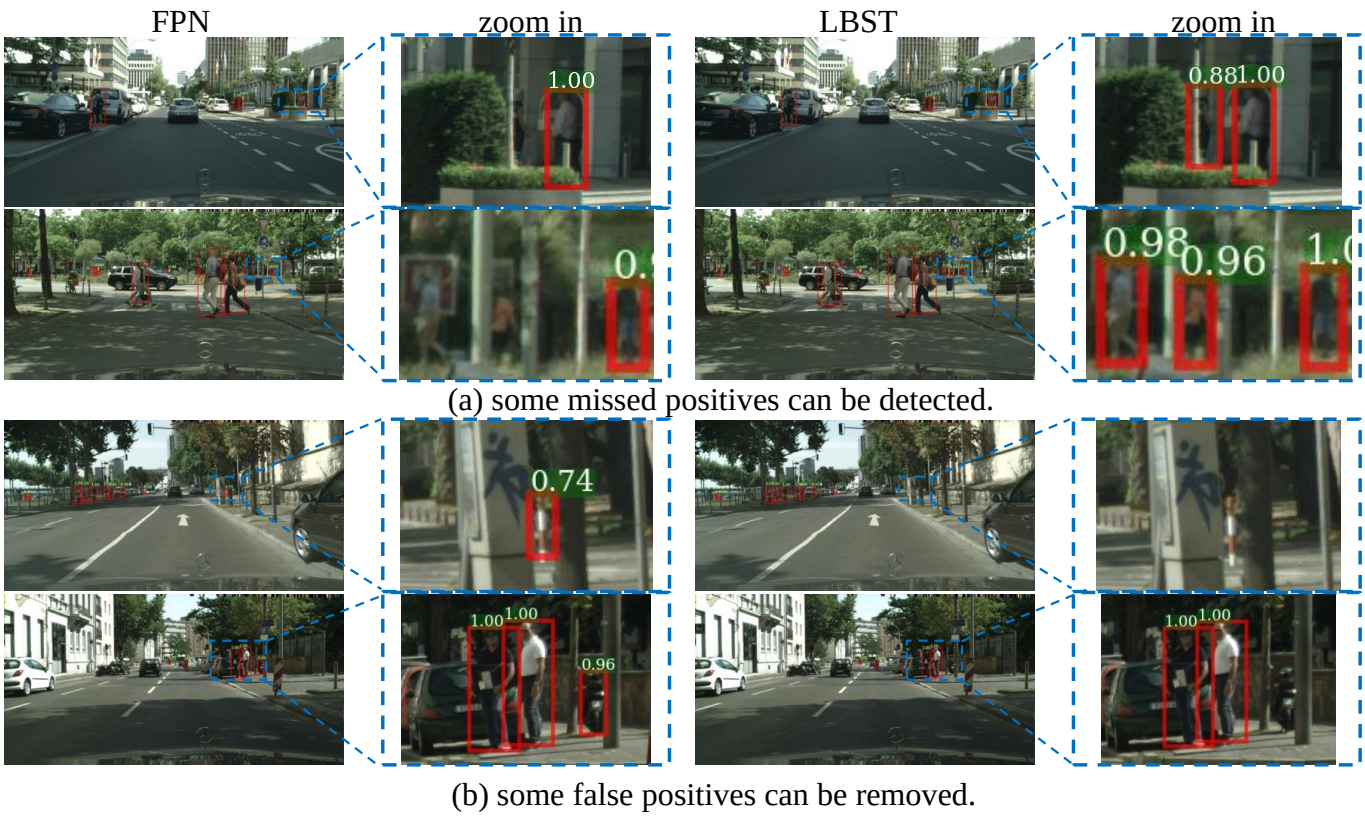
method	training pedestrians	R	RS	HO	R+HO	A
FPN [32]	*h50o5*: *h*>50, *occ*<0.5	14.0	20.4	50.2	31.0	42.9
+LB	*h50o05*: *h*>50, *occ*<0.5	13.0	19.7	49.3	29.8	42.0
+ST	*h50o05*: *h*>50, *occ*<0.5	13.7	19.6	48.9	30.2	42.1
LBST	*h50o05*: *h*>50, *occ*<0.5	12.6	18.6	48.7	29.1	41.5
FPN [32]	*h30o05*: *h*>30, *occ*<0.5	14.9	19.6	50.3	31.5	40.8
+LB	*h30o05*: *h*>30, *occ*<0.5	14.0	19.1	49.3	30.4	39.8
+ST	*h30o05*: *h*>30, *occ*<0.5	14.3	18.7	49.0	30.2	40.1
LBST	*h30o05*: *h*>30, *occ*<0.5	13.6	18.6	48.2	29.7	38.8

Evaluation: For performance evaluation on the Cityscapes [54] and Caltech [15] pedestrian datasets, the log-averaged miss rate under FPPI=[0.01,1] is used. FPPI means false positive per image.

B. Experiments on the Citypersons dataset

Settings: The backbone is the famous ResNet50 [20] which is pre-trained on ImageNet [12]. The total number of iterations is $30k$. The initial learning rate is 0.003. After that, it decreases by a factor of 10 at the $20k$ and $25k$ iterations. At the test stage, four different evaluation metrics are both used. Specifically, the `reasonable` (50-∞ pixels tall and 0-0.35 occ), `small` (50-75 pixels tall and 0-0.35 occ), `heavy` (50-∞ pixels tall and 0.35-0.8 occ), and `all` (20-∞ pixels tall and 0-0.8 occ) sets are used.

Ablation: Table I demonstrates the effectiveness of the proposed two modules (i.e., location bootstrap and semantic transition). The baseline FPN [32] uses the 1×1 convolution as the skip-layer connection. At the training stage, the pedestrians at least 50 pixels tall and less than 50% occlusion (*h50o05*) and pedestrians at least 30 pixels tall and less than 50% occlusion (*h30o05*) are respectively used. Based on Table I, it



(a) some missed positives can be detected.

(b) some false positives can be removed.

Fig. 5. Detection results of FPN and the proposed LBST. It can be seen that the proposed method can not only detect some missed positives, but also remove some false positives.

TABLE II
RESULTS BY TRAINING AND TESTING ON ALMOST THE ALL DATA.

FPN	+LB	+ST	LBST
44.3	43.5	43.6	43.2

TABLE III
RESULTS OF DIFFERENT LOCATION BOOTSTRAP SCHEMES.

the baseline	the fixed schme	the averaged scheme
31.0	30.1	29.8

is concluded that (1) Both location bootstrap (LB) module and semantic transition module (ST) can improve detection performance for small-scale pedestrians and occluded pedestrians. For example, using *h30o05*, the proposed method with LB outperforms the baseline (FPN) by 0.5% and 1.0% on **RS** and **HO**; the proposed method with ST outperforms FPN by 0.9% and 1.3% on **RS** and **HO**. (2) When integrating LB and ST together, the proposed method can further improve detection performance. For example, using *h50o05* (or *h30o05*), the proposed LBST outperforms FPN by 1.9% (or 1.8%) on **R+HO**. (3) We also calculate the performance on **RS+HO**. Using *h50o05*, the miss rates of FPN and LBST are 40.9% and 39.2%. Namely, LBST outperforms FPN by 1.7%.

By using almost all available data in the training set and the test set, Table II further compares the miss rates of different methods. It can be seen that our proposed methods also outperform FPN.

The weights in location bootstrap are calculated by Eq. 3. For simplicity, it is called as the averaged scheme. Another simple and direct way is using a linear and fixed scheme (i.e., $w_i = 2.5 - 2 * O_i$). For simplicity, it is as the fixed scheme. Table III compares the two schemes on **R+HO** by using $h50o05$. It is seen that the two schemes are both better than the baseline FPN and the averaged scheme is a little better than the fixed scheme. Thus, the averaged scheme is used.

Parameter settings There are some hyper parameters in our methods. (1) *the size of large kernel.* When $k = 3, 5, 7, 9$, the miss rates on **R+HO** is 30.8%, 30.6%, 30.2%, and 30.3%. Thus, $k = 7$ is used. (2) *multiple branches of different large kernel.* When using three branches with the large kernel sizes of 3,5,7, the miss rate on **R+HO** is 30.0%, which is a little improvement. For simplicity, we do not use multiple branches of different large kernel. (3) α *in location bootstrap.* When $\alpha = 1.0, 2.0, 3.0$, the miss rates on **R+HO** is 30.4%, 29.8%, and 30.0%. Thus, $\alpha = 2$ is used.

Qualitative results: Fig. 5 shows some qualitative detection results of FPN [32] and the proposed LBST by trained on $h50o05$. Fig. 5(a) shows two examples that the proposed LBST can detect some missed small-scale pedestrians and occluded pedestrians. For example, in the second row, LBST can detect

TABLE IV
COMPARISON WITH SOME STATE-OF-THE-ART METHODS ON THE CITYPERSONS VAL SET. h MEANS THE PEDESTRIAN HEIGHT, *occ* MEANS THE PEDESTRIAN OCCLUSION RATIO. **R** MEANS THE REASONABLE SET, **RS** MEANS THE SMALL SET, **HO** MEANS THE HEAVY SET, **R+HO** MEANS THE REASONABLE+HEAVY SET, AND **A** MEANS THE ALL SET.

method	scale	training pedestrians	R	RS	HO	R+HO	A
FRCNN+ATT [57]	×1.0	*h50o035*: h>50, *occ*<0.35	15.9	-	56.6	38.2	-
RepLoss [49]	×1.0	*h50o035*: h>50, *occ*<0.35	13.2	22.3	56.9	31.8	44.5
ORCNN [56]	×1.0	*h25o05*: h>25, *occ*<0.5	12.8	-	55.7	-	-
ALFNet [36]	×1.0	*h50o1*: h>50, *occ*<1.0	**12.0**	-	51.9	-	-
TTL(MRF) [46]	×1.0	-	14.4	-	52.0	-	-
LBST	×1.0	*h50o035*: h>50, *occ*<0.35	12.8	**18.8**	53.7	30.6	43.2
LBST	×1.0	*h50o05*: h>50, *occ*<0.5	12.6	**18.6**	48.7	29.1	41.5
LBST	×1.0	*h50o07*: h>50, *occ*<0.7	13.4	19.6	**42.0**	**27.9**	40.7
LBST	×1.0	*h30o05*: h>30, *occ*<0.5	13.6	18.7	48.2	29.7	38.8
LBST	×1.0	*h30o07*: h>30, *occ*<0.7	13.3	19.5	43.7	28.0	**38.1**
AdaptedRCNN [54]	×1.3	*h50o035*: h>50, *occ*<0.35	12.8	-	-	-	-
RepLoss [49]	×1.3	*h50o035*: h>50, *occ*<0.35	11.6	-	55.3	-	-
ORCNN [56]	×1.3	*h25o05*: h>25, *occ*<0.5	**11.0**	**13.0**	51.9	29.4	39.4
PDOE+RPN [60]	×1.3	*h50o07*: h>50, *occ*<0.7	11.2	47.4	44.2	-	43.4
LBST	×1.3	*h50o035*: h>50, *occ*<0.35	11.3	**15.0**	50.5	28.8	40.8
LBST	×1.3	*h50o05*: h>50, *occ*<0.5	11.4	14.3	45.2	27.1	39.3
LBST	×1.3	*h50o07*: h>50, *occ*<0.7	11.2	15.9	**38.9**	**24.8**	37.7
LBST	×1.3	*h30o05*: h>30, *occ*<0.5	11.3	14.1	42.9	26.9	35.3
LBST	×1.3	*h30o07*: h>30, *occ*<0.7	11.4	15.4	39.9	25.6	**34.6**

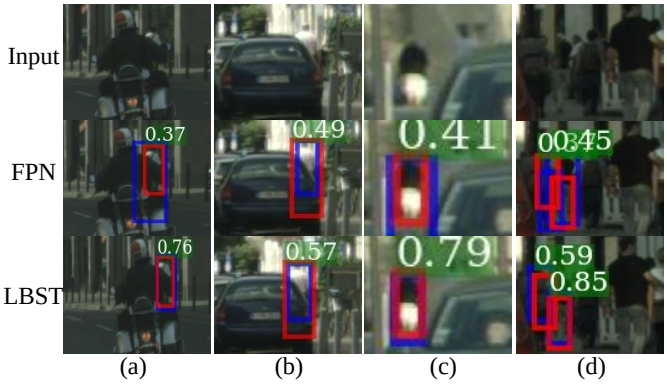


Fig. 6. Bounding locations of FPN and the proposed method, where the number is the IoU between ground-truth (red) and bounding boxes (blue). (a) The first row is image samples. (b) The second row is the location of baseline. (c) The third row is the location of the proposed method.

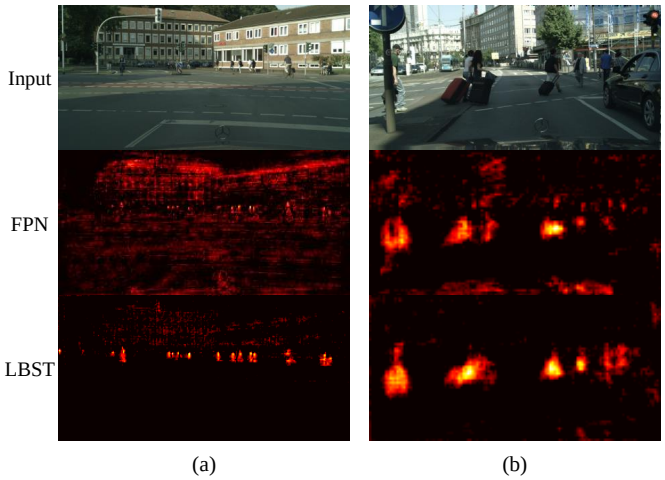


Fig. 7. Feature visualization of FPN [32] and the proposed methods. (a) The first row is the input images. (b) The second row is the feature maps of baseline. (c) The third row is the feature maps of the proposed method.

two small-scale pedestrians which are missed by FPN. Fig. 5(b) shows two examples that the proposed LBST can remove some false positives. For example, in the second row, our LBST can remove the false positive which are mistakenly recognized as a pedestrian by FPN.

Accurate location: Moreover, Fig. 6 further shows that some pedestrians can be accurately detected by the proposed LBST but not accurately located by FPN [32]. The first row is the input images, the second row is the location results of FPN, and the third row is the location results of the proposed LBST. For example, in the first column, the occluded pedestrian detected by FPN has 37% overlap with the ground-truth, while it detected by the proposed LBST has 76% overlap with the ground-truth.

Feature Visualization: To better demonstrate why the proposed method can improve detection performance, Fig. 7 visualizes the features of FPN [32] and the proposed LBST. The feature map which has the largest response on the pedestrians is chosen. The first row is the input images, the second row is the feature maps of baseline, and the third row is the feature maps of the proposed LBST. Compared with the feature maps of FPN, the feature maps of proposed LBST are less noisy. The reason is that semantic transition module contains more contextual information and enhances the feature semantic. As a result, the features are more robust.

Comparisons with others: Finally, Table IV compares

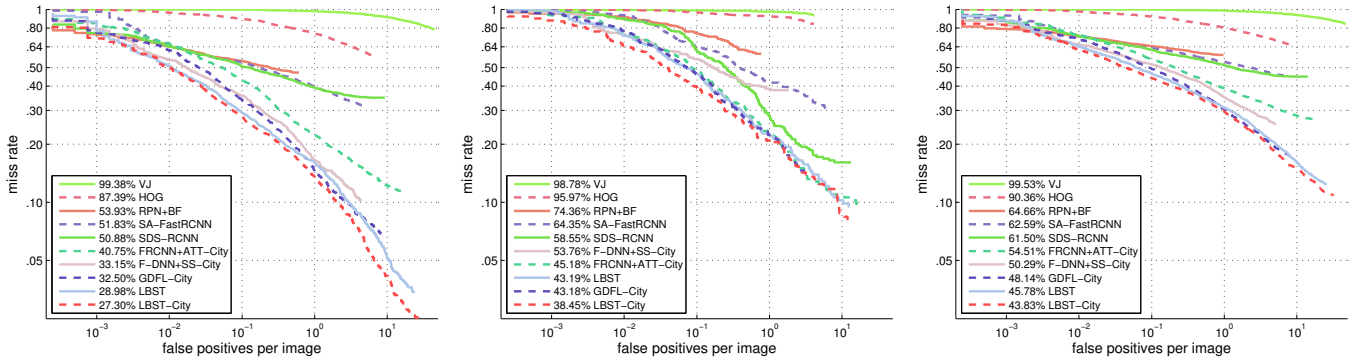


Fig. 8. Results on the Caltech test set. For left to right, miss rates on the `medium` (**M**), `heavy` (**HO**), and `all` (**A**) sets are shown. '-City' means that the extra Citypersons dataset is used for training.

the proposed LBST with some state-of-the-art methods (i.e., AdaptedRCNN [54], FRCNN+ATT [57], TTL [46], RepLoss [49], ORCNN [56], and ALFNet [36]) on the Citypersons `val` set. Because these state-of-the-art methods do not use the same settings for choosing the positive pedestrians at the training stage, the detection performance of the proposed method at the five different subsets are given. Namely, the pedestrians of h>50, occ<0.35 (called *h50o035*), the pedestrians of h>50, occ<0.5 (called *h50o05*), the pedestrians of h>50, occ<0.7 (called *h50o07*), the pedestrians of h>30, occ<0.5 called *h30o05*, and the pedestrians of h>30, occ<0.7 (called *h30o07*) are used for training. It can be seen that (1) The proposed methods achieve best performance on the **HO**, **R+HO**, and **A** with the similar settings. For example, with similar settings at the scale ×1.0, LBST outperforms RepLoss by 0.4% and 3.2% on **R** and **HO**. With the similar setting at scale ×1.3, LBST outperforms ORCNN by 4.8% on **HO** and outperforms PDOE+RPN by 5.3% on **A**. (2) On the **R** and **RS**, the proposed method almost achieves state-of-the-art performance with the similar settings, which is slightly inferior to ORCNN. Please note that ORCNN uses the smaller-scale training pedestrians (h>25) and dense anchor settings [58].

Effect of the training pedestrians: The state-of-the-art methods use the different settings to choose the positive pedestrians during training and do not discuss the effect of different settings. To help us better understand their effects on final detection performance, it is necessary and useful to discuss their effects. Based on the results of proposed LBST which uses five different settings (i.e., *h50o035*, *h50o05*, *h50o07*, *h30o05*, and *h30o07*), it can be concluded that: (1) *On occluded pedestrian detection*, more heavily occluded pedestrians for training have the positive effect. For example, using ×1.0 input image, LBST with $h50o07$ outperforms LBST with $h50o05$ by 6.7% on **HO**. (2) *On small-scale pedestrian detection*, more small-scale pedestrians for training have positive effect. For example, using ×1.0 input image, LBST with $h30o05$ outperforms LBST with $h50o05$ by 2.7% on **A**, while LBST with $h30o05$ has the similar performance as that with $h50o05$ on **R+HO**. It means that the improvement is from small-scale pedestrians of 20-50 pixels tall. (3) *On standard pedestrian detection* (`reasonable`), more occluded pedestrians and small-scale pedestrians have little effect.

TABLE V
MISS RATES ON THE CALTECH `TEST` SET. **R** MEANS THE `REASONABLE` SET, **M** MEANS THE `MEDIUM` SET, **HO** MEANS THE `HEAVY` SET, **R+HO** MEANS THE `REASONABLE+HEAVY` SET, AND **A** MEANS THE `ALL` SET. "+CITY" (OR "+COCO") MEAN THAT THE METHOD IS LEARNED BASED ON BOTH THE CALTECH AND CITYPERSONS (OR COCO [33]) DATASETS.

method	data	R	M	HO	R+HO	A
RPN+BF [53]	Caltech	9.58	53.93	74.36	24.01	64.66
SA-RCNN [23]	Caltech	9.68	51.83	64.35	21.92	62.59
SDS-RCNN [4]	Caltech	**7.36**	50.88	58.55	19.72	61.50
PDOE+RPN [60]	Caltech	7.6	-	44.4	-	-
LBST	Caltech	9.26	**28.98**	43.19	17.02	**45.78**
F-DNN-SS [16]	+City	8.18	33.15	53.76	18.82	50.29
FRCNN-ATT [57]	+City	10.33	40.75	45.18	18.21	54.51
GDFL [27]	+City+COCO	**7.85**	32.50	43.18	15.64	48.14
LBST	+City	8.59	**27.30**	**38.45**	**15.39**	**43.83**

C. Experiments on the Caltech pedestrian dataset

Settings: Experiments are further conducted on the Caltech dataset [15]. The images are twice upsampled at the training and test stages. The total number of iterations is $80k$. The initial learning rate is 0.001. After that, it decreases at the $60k$ iterations by a factor of 10. The standard evaluation metric on the Caltech dataset [15] is used. Specifically, the `reasonable` (50-∞ pixels tall and 0-0.35 occ), `medium` (30-80 pixels tall and 0-0.35 occ), `heavy` (50-∞ pixels tall and 0.35-0.8 occ), and `all` (20-∞ pixels tall and 0-0.8 occ) sets are used.

Caltech: The top of Table V shows miss rates of these methods only trained on the Caltech. It can be concluded that: (1) The proposed LBST has best performance on the small-scale pedestrian detection (i.e.,**M**) and occluded pedestrian detection (i.e., **HO**). LBST outperforms SDS-RCNN [4] by 11.90% on **M** and 15.36% on **HO**. (2) Though LBST does not achieve best performance on **R**, it outperforms all the other methods on the **R+HO** and **A** sets. LBST outperforms SDS-RCNN by 2.70% on **R+HO**.

Caltech+: The bottom of Table V shows miss rates of these methods which are firstly trained on the extra dataset (e.g., COCO [33] or Citypersons [54]) and then fine-tuned on the Caltech. With the extra Citypersons [54], LBST outperforms

all the other methods on **M**, **HO**, **HO**, and **A**. For example, LBST outperforms GDFL [27] by 5.20% on **M**, 4.73% on **HO**, 0.25% on **R+HO**, and 4.31% on **A**. Fig. 8 further plots the curves of these methods on **M**, **HO**, and **A**. The proposed LBST steadily outperforms other methods.

V. Conclusion

In this paper, we realized that many small-scale pedestrians and occluded pedestrians can be found but not well located. To solve this problem, two simple modules (i.e., location bootstrap and semantic transition module) are proposed for pedestrian detection. The location bootstrap re-weights the loss of predicted bounding boxes, while the semantic transformation module enhances the feature semantic and adds more contextual information for skip-layer fusion. Experiments on the challenging Citypersons and Caltech pedestrian datasets demonstrate that the proposed methods can improve detection accuracy for small-scale pedestrians and occluded pedestrians.

References

[1] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, Attention CoupleNet: Fully convolutional attention coupling network for object detection, *IEEE Trans. Image Processing*, vol. 28, no. 1, pp. 113-126, 2019. 2

[2] T. Chen, L. Lin, X. Wu, N. Xiao, and X. Luo, Learning to segment object candidates via recursive neural networks, *IEEE Trans. Image Processing*, vol. 27, no. 12, pp. 5827-5839, 2018. 2

[3] X. Zhang, L. Cheng, B. Li, and H.-M. Hu, Too Far to see? not really!pedestrian detection with scale-aware localization policy, *IEEE Trans. Image Processing*, vol. 27, no. 8, pp. 3703-3715, 2018. 2

[4] G. Brazil, X. Yin, and X. Liu, Illuminating pedestrians via simultaneous detection & segmentation, *Proc. IEEE International Conference on Computer Vision*, 2017. 2, 8

[5] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, *Proc. European Conference on Computer Vision*, 2016. 1, 2

[6] Z. Cai and N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[7] J. Cao, Y. Pang, and X. Li, Pedestrian detection inspired by appearance constancy and shape symmetry, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[8] J. Cao, Y. Pang, and X. Li, Learning multilayer channel features for pedestrian detection, *IEEE Trans. Image Processing*, vol. 26, no. 7, pp. 3210–3220, 2017. 2

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 2

[10] Z. Chen, S. Huang, and D. Tao, Context refinement for object detection, *Proc. European Conference on Computer Vision*, 2018. 2

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, The cityscapes dataset for semantic urban scene understanding, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[13] P. Dollár, R. Appel, S. Belongie, and P. Perona, Fastest feature pyramids for object detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2016. 2

[14] P. Dollár, Z. Tu, P. Perona, and S. Belongie, Integral channel features, *Proc. British Machine Vision Conference*, 2009. 2

[15] P. Dollár, C. Wojek, B. Schiele, and P. Perona, Pedestrian detection: An evaluation of the state of the art, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012. 2, 5, 8

[16] X. Du, M. El-Khamy, J. Lee, and L. Davis, Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection, *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2017. 8

[17] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, DSSD : Deconvolutional single shot detector, *arXiv:1701.06659*, 2017. 2

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, *Proc. Advances in Neural Information Processing Systems*, 2014. 3

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn, *Proc. IEEE International Conference on Computer Vision*, 2017. 5

[21] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *Proc. IEEE International Conference on Computer Vision*, 2016. 3

[22] J. Hosang, M. Omran, R. Benenson, and B. Schiele, Taking a deeper look at pedestrians, *Proc. IEEE International Conference on Computer Vision*, 2015. 1

[23] X. L. J. Li, S. Shen, T. Xu, and S. Yan, Scale-aware fast r-cnn for pedestrian detection, *IEEE Trans. Multimedia*, vol. 20, no.4, pp. 985-996, 2017. 1, 2, 8

[24] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, Acquisition of localization confidence for accurate object detection, *Proc. European Conference on Computer Vision*, 2018. 3

[25] Y. Kim, B.-N. Kang, and D. Kim, San: Learning relationship between convolutional features for multi-scale object detection, *Proc. European Conference on Computer Vision*, 2018. 1, 3

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Proc. Advances in Neural Information Processing Systems*, 2012. 1, 2

[27] C. Li, J. Lu, G. Wang, and J. Zhou, Graininess-aware deep feature learning for pedestrian detection, *Proc. European Conference on Computer Vision*, 2018. 2, 8, 9

[28] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, Perceptual generative adversarial networks for small object detection, *Proc. IEEE International Conference on Computer Vision*, 2017. 1, 3

[29] B. Li, T. Wu, L. Zhang, and R. Chu, Auto-Context R-CNN, *Proc. European Conference on Computer Vision*, 2018. 2

[30] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, Attentive contexts for object detection, *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, 2017. 2

[31] C. Lin, J. Lu, G. Wang, and J. Zhou, Graininess-aware deep feature learning for pedestrian detection, *Proc. European Conference on Computer Vision*, 2018. 2

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 4, 5, 6, 7

[33] T.-Y. Lin, M. Maire, S. Belongie, P. P. J. Hays, P. D. D. Ramanan, and C. L. Zitnick, Microsoft coco: Common objects in context, *Proc. European Conference on Computer Vision*, 2014. 8

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal Loss for Dense Object Detection, *Proc. IEEE International Conference on Computer Vision*, 2017. 4

[35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, SSD: Single shot multibox detector, *Proc. European Conference on Computer Vision*, 2014. 2

[36] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, Learning efficient single-stage pedestrian detectors by asymptotic localization fitting, *Proc. European Conference on Computer Vision*, 2016. 2, 7, 8

[37] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, What can help pedestrian detection? *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[38] W. Nam, P. Dollár, and J. H. Han, Local decorrelation for improved pedestrian detection, *Proc. Advances in Neural Information Processing Systems*, 2014. 2

[39] J. Noh, S. Lee, B. Kim, and G. Kim, Improving occlusion and hard negative handling for single-stage pedestrian detectors, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[40] Y. Pang, J. Cao, J. Wang, and J. Han, JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images, *IEEE Trans. Information Forensics and Security*, vol. 14, no. 12, pp. 3322–3331, 2019. 1

[41] C. Peng, and X. Zhang, G. Yu, and G. Luo, and J. Sun, Large kernel matters Improve semantic segmentation by global convolutional network, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017. 2

[42] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement , *arXiv:1804.02767*, 2018. 2

[43] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *Proc. Advances in Neural Information Processing Systems*, 2015. 1, 2
[44] A. Shrivastava, A. Gupta, R. Girshick, Training Region-based Object Detectors with Online Hard Example Mining, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4
[45] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv:1409.1556*, 2014. 1, 2, 3
[46] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation, *Proc. European Conference on Computer Vision*, 2018. 3, 7, 8
[47] Y. Tian, P. Luo, X. Wang, and X. Tang, Deep learning strong parts for pedestrian detection, *Proc. IEEE International Conference on Computer Vision*, 2015. 1, 3
[48] S. Wang, J. Cheng, H. Liu, and M. Tang, PCN: Part and context information for pedestrian detection with CNNs, *Proc. British Machine Vision Conference*, 2018. 2
[49] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, Repulsion loss: Detecting pedestrian in a crowd, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 7, 8
[50] B. Yang, J. Yan, Z. Lei, and S. Z. Li, Convolutional channel features, *Proc. IEEE International Conference on Computer Vision*, 2015. 2
[51] F. Yang, W. Choi, and Y. Lin, Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2
[52] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, H. Zhou, and X. Wang, Crafting GBD-Net for object detection, *Proc. European Conference on Computer Vision*, 2016. 2
[53] L. Zhang, L. Lin, X. Liang, and K. He, Is faster r-cnn doing well for pedestrian detection, *Proc. European Conference on Computer Vision*, 2016. 2, 8
[54] S. Zhang, R. . Benenson, and B. Schiele, Citypersons: A diverse dataset for pedestrian detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 5, 7, 8
[55] S. Zhang, R. Benenson, and B. Schiele, Filtered channel features for pedestrian detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2
[56] S. Zhang, L. Wen, X. Bian, Z. Lei, , and S. Z. Li, Occlusion-aware r-cnn: Detecting pedestrians in a crowd, *Proc. European Conference on Computer Vision*, 2018. 1, 3, 7, 8
[57] S. Zhang, J. Yang, and B. Schiele, Occluded pedestrian detection through guided attention in cnns, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3, 7, 8
[58] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, S3fd: Single shot scale- invariant face detector, *Proc. IEEE International Conference on Computer Vision*, 2017. 8
[59] C. Zhou and J. Yuan, Multi-label learning of part detectors for heavily occluded pedestrian detection, *Proc. IEEE International Conference on Computer Vision*, 2017. 2
[60] C. Zhou and J. Yuan, Bi-box regression for pedestrian detection and occlusion estimation, *Proc. European Conference on Computer Vision*, 2018. 3, 7, 8